Supporting Information

Keele *et al.* 10.1073/pnas.0802203105

SI Text

Study Subjects. Plasma samples were obtained from 102 subjects with acute or very recent HIV-1 subtype B infection. These included embargoed serial collections from 54 source plasma donors (ZeptoMetrix, Inc.; SeraCare Life Sciences, Inc.) who became HIV-1-infected during prolonged periods of once- or twice-weekly plasma donations, and 48 other patients who presented to U.S. or Trinidad (3) healthcare facilities generally with symptoms of recently acquired sexually transmitted diseases or acute HIV-1 infection. Demographic information is provided in Dataset S1, Dataset S2, Dataset S3, and Dataset S4. Plasma samples from 43 subjects from the U.S. or Trinidad with chronic, treatment-naïve HIV-1 infection were obtained as controls (Dataset S6). All subjects gave informed consent, and plasma collections were performed with institutional review board and other regulatory approvals. Blood specimens were generally collected in acid citrate dextrose and plasma separated and stored at −20°C to −70°C.

Laboratory Staging. Plasma samples were tested for HIV-1 RNA, p24 antigen, and viral-specific antibodies by a battery of commercial tests. These included quantitative bDNA 3.0 (Chiron) or Amplicor vRNA (Roche) assays; Coulter or Roche p24 Ag assays; Anti-HIV-1/2 Plus O (Genetic Systems) and AntiHIV-1/2 3rd Generation (Abbott) EIAs; and HIV-1 Western Blot Kit (Genetic Systems). Based on these test results, subjects were staged according to the Fiebig laboratory classification system for acute and early HIV-1 infection (4). The duration of the eclipse phase (before the detection of plasma viral RNA) was estimated to be 10 days (range, 7–21 days) (5–10).

Viral RNA Extraction and cDNA Synthesis. For each sample, ≈20,000 viral RNA copies were extracted by using the QIAamp Viral RNA Mini Kit (Qiagen). For subjects in whom sequential samples were analyzed beginning in the early ramp-up phases of infection (Fiebig stage I), samples containing as few as 10 vRNA molecules were processed and analyzed. RNA was eluted and immediately subjected to cDNA synthesis. Reverse transcription of RNA to single-stranded cDNA was performed with SuperScript III reverse transcriptase by using methods recommended by the manufacturer (Invitrogen Life Technologies). Briefly, each cDNA reaction included 1× reverse transcription (RT) buffer, 0.5 mM each deoxynucleoside triphosphate, 5 mM DTT, 2 units/μl RNaseOUT (RNase inhibitor), 10 units/μl SuperScript III reverse transcriptase, and 0.25 μM antisense primer Env3out 5′-TTGCTACTTGTGATTGCTCCATGT-3′ (nt 8913–8936 HXB2). In some experiments, a different antisense primer OFM19 [5′- GCACTCAAGGCAAGCTTTATTGAGGCTTA-3′ (nt 9604 to 9632 HXB2)] was used. The mixture was incubated at 50°C for 60 min, followed by an increase in temperature to 55°C for an additional 60 min. The reaction was then heat-inactivated at 70°C for 15 min and then treated with RNaseH at 37°C for 20 min. The newly synthesized cDNA was used immediately or kept frozen at −80°C.

Single Genome Amplification. cDNA was serially diluted and distributed among wells of replicate 96-well plates so as to identify a dilution where PCR positive wells constituted less than 30% of the total number of reactions. At this dilution, most wells contain amplicons derived from a single cDNA molecule. This was confirmed in every positive well by direct sequencing of the amplicon and inspection of the sequence for mixed bases (double peaks), which would be evidence of priming from more than one original template or the introduction of PCR error in early cycles. Any sequence with evidence of mixed bases was excluded from further analysis. PCR amplification was carried out in the presence of 1× High Fidelity Platinum PCR buffer, 2 mM $MgSO_4$, 0.2 mM each deoxynucleoside triphosphate, 0.2 μM each primer, and 0.025 units/μl Platinum Taq High Fidelity polymerase in a 20-μl reaction (Invitrogen). First-round PCR primers included sense primer Env5out 5′-TAGAGCCCTGGAAGCATCCAGGAAG-3′ (nt 5853–5877 HXB2) and antisense primer Env3out 5′-TTGCTACTTGTGATTGCTCCATGT-3′ (nt 8913–8936 HXB2), which generated an ≈3-kb amplicon. PCR was performed in MicroAmp 96-well reaction plates (Applied Biosystems) with the following PCR parameters: 1 cycle of 94°C for 2 min; 35 cycles of a denaturing step of 94°C for 15 s, an annealing step of 55°C for 30 s, an extension step of 68°C for 4 min, followed by a final extension of 68°C for 10 min. Next, 2 μl from first-round PCR product was added to a second-round PCR that included the sense primer Env5in 5′-caccTTAGGCATCTCCTATGGCAGGAAGAAG-3′ (nt 5957–5983 HXB2) and antisense primer Env3in 5′-GTCTCGAGATACTGCTCCCACCC-3′ (nt 8904–8882 HXB2). The addition of "cacc" to the sense primer allowed for directional cloning of the amplicon (see Env gene cloning section below). The second-round PCR was carried out under the same conditions used for first-round PCR but with a total of 45 cycles. Amplicons were inspected on precasted 1% agarose E-gels 96 (Invitrogen Life Technologies). All PCR procedures were carried out under PCR clean room conditions with procedural safeguards against sample contamination, including prealiquoting of all reagents, use of dedicated equipment, and physical separation of sample processing from pre- and post-PCR amplification steps.

DNA Sequencing. *Env* gene amplicons were directly sequenced by cycle-sequencing by using BigDye terminator chemistry and protocols recommended by the manufacturer (Applied Biosystems). Sequencing reaction products were analyzed with an ABI 3730xl genetic analyzer (Applied Biosystems). Both DNA strands were sequenced by using partially overlapping fragments. Individual sequence fragments for each amplicon were assembled and edited by using the Sequencher program 4.7 (Gene Codes). Inspection of individual chromatograms allowed for the identification of amplicons derived from single versus multiple templates. The absence of mixed bases at each nucleotide position throughout the entire *env* gene was taken as evidence of single genome amplification from a single viral RNA/cDNA template. This quality control measure enabled us to exclude from the analysis amplicons that resulted from PCR-generated *in vitro* recombination events or *Taq* polymerase errors and to obtain multiple individual *env* sequences that proportionately represented those circulating *in vivo* in HIV-1 virions.

Sequence Alignments. All alignments were initially made with GeneCutter (www.hiv.lanl.gov) to compensate for frame-shifting mutations. Because the alignment was large and the *env* genes interspersed with insertions and deletions, and because automatic multiple sequence alignment programs are often not effective in hypervariable regions, we developed an iterative alignment process to hand-check and improve the alignments. We, thus, generated a consensus sequence for the sequence set from each individual, which we then extracted from the full

alignment and hand adjusted to improve the alignment. The within-patient sets were then realigned to each patient consensus, each within-patient alignment was again hand-adjusted, and a new consensus for each patient was generated. This process was iterated several times to improve the alignments. To generate the final consensus sequence for each patient, ties near regions of insertion and deletions were resolved by considering the proximal codons and context. The full alignment is available in a supplemental data file, and the sequences are also available through GenBank. All 4,357 *env* sequences from acute and chronic patients were deposited in GenBank, and edited envelope alignments can be accessed at www.hiv.lanl.gov/content/sequence/hiv/user_alignments/keele.

***Env* Diversity Analysis.** We classified two very distinctive levels of within-patient *env* diversity that we observed in the 102 study subjects as either "homogeneous" or "heterogeneous." This was done by using three different strategies that all concurred. First, we visually inspected the samples by using neighbor-joining phylogenies and the *Highlighter* tool (www.hiv.lanl.gov) and found that 21 samples clearly had much greater diversity than 81 others. We then looked at all pairwise Hamming distances (HD) (defined as the number of base positions at which the two genomes differ, excluding gaps) within each sample. The same 21 heterogeneous samples exhibited distinct peaks with a multimodal distribution inconsistent with expansion from a single infecting virus. Last, to formalize the criteria and test whether the 21 heterogeneous samples reflected transmission of multiple variants, we used the mathematical model described below to predict the expected maximum HD that could be observed under a homogeneous infection assumption (i.e., infection by a single virus), given the Fiebig stage of the sample. If the maximum HD in the sample was much greater than expected, the observed diversity was considered to have originated at a time before transmission, i.e., in the donor, indicating that multiple strains transmitted from the donor to the recipient established the infection; this was, again, the case for all 21 heterogeneous samples. For the homogeneous samples, we considered the possibility that these individuals had been infected by a single virus (or infected cell) or by two very closely related viruses. Either scenario could result in a low overall *env* diversity, but in the case of transmission of two very closely related viruses, the distribution of HDs would not fit model expectations. We found this to be the case in three of the 81 subjects with homogeneous infections.

Star Phylogeny. With no selection pressure, one can expect homogeneous viral populations to evolve from a founder strain in a star-like phylogeny, i.e., all evolving sequences coalesce at the founder. The veracity of this proposition can be investigated by inspecting the sequence alignment. Because mutations are rare, one does not expect shared mutations in a star phylogeny. When this is indeed the case, the distribution of intersequence HDs is constrained to be a self-convolution (defined below) of the distribution of the HDs from the sequences to the ancestral sequence. In particular, for every pair of sequences s_1 and s_2, let $HD[s_1,s_2]$ be the number of base positions at which the two differ and the probability distribution it follows be $P_I(HD)$. Next, we compare each sequence in the sample with the consensus sequence (which we assume to be the founder strain) and compute the corresponding HD distribution. Denoting s_0 the founder strain, for every sequence s_1 we compute $HD[s_0,s_1]$ and we denote $P_C(HD)$ the distribution it follows. Then, under a star-phylogeny evolution, $P_I(HD)$ is given by the self-convolution of $P_C(HD)$:

$$P_I(\mathrm{HD}=n)=\sum_{k=0}^{n} P_C(\mathrm{HD}=k)P_C(\mathrm{HD}=n-k) \quad [1]$$

Occasional deviations from a star phylogeny are, however, expected. The sampling of 30 sequences, for example, from a later generation of an exponentially growing population with 6-fold growth per generation has an ≈5% chance of including a pair of sequences that shares five initial generations, has a 25% chance of those sharing the first four, and is overwhelmingly likely to include sequences that share three ancestors. However, because the rate of mutations in the region under study is approximately one per 20 generations (see next section), this leads to only an ≈10% chance of finding sequences sharing a pair of mutations, and a <1% chance of sharing more than that. The probabilities are slightly enhanced by the early stochastic events that can lead to the virus producing less than six descendants in some generations, but it remains overwhelmingly likely that the sequences share few mutations. Thus, when a sample had two or more sublineages of sequences that were defined by more than two shared mutations, and the sample was classified as Fiebig stage II (and so before immune selection), the observation is best explained by transmission of multiple closely related viruses (three such cases were identified). In later Fiebig stages, early CTL-driven immune selection may contribute to such a pattern and selection cannot be distinguished from transmission of multiple viruses.

Mathematical Model. We assume a homogeneous infection in which the virus grows exponentially with no selection pressure, no recombination, no occurrence of back mutations and a constant mutation rate across positions and across lineages. Under this scenario, the HD frequency distribution is given by a Poisson distribution whose mean depends linearly on the number of generations since the founder strain. We used previously estimated parameters of HIV-1 generation time (2 days) (11), reproductive ratio (R_0, 6) (12), and reverse transcriptase point mutation rate ($\varepsilon = 2.16 \times 10^{-5}$) (13) and assumed that the initial virus replicated exponentially by infecting exactly R_0 new cells at each generation, which, for simplicity, we assumed happened in two equal bursts at τ and 2τ. The reverse transcriptase error rate estimate (13) is based on sequencing virus produced *in vitro* after a single round of replication. If a mutation occurs that is lethal with regard to viral production, it would not be detected in this assay, and such mutations may be similarly reduced in the natural, *in vivo* situation. On the other hand, lethal mutations that were not infectious would be retained in the single round of replication assay but may be selected against *in vivo*; hence, the mutation rate we are using in the model will have a bias toward being greater than the substitution rate we observe *in vivo*, potentially resulting in slight underestimates of the time to the MRCA.

The intersequence HDs are not independent, but, because of the star phylogeny, they are the pairwise sums of a set of independent Poisson distributed variates. The form of their distribution, including the (singular) covariance matrix, is, therefore, known up to one unknown parameter, the λ of the underlying Poisson distribution. We estimated this parameter by fitting the observed data to the expected form by using a maximum likelihood method and assessed the goodness of fit by using a χ^2 goodness-of-fit test statistic calculated from a singular value decomposition of the covariance matrix. When the data were consistent with a Poisson distribution, we used the λ of the best fitting distribution to estimate a divergence time from the most recent common ancestor (MRCA) based on the estimated number of generations required to achieve the observed distribution. One can, in fact, show the following relationship for λ:

$$\lambda(t) = \varepsilon N_B\left(\frac{5}{8}t\,\frac{1+\varphi}{\varphi} + \frac{1-\varphi}{\varphi^2}\right) \quad [2]$$

Therefore, once we obtain a best fitting Poisson distribution, we calculate its mean λ^* and use the above time–dependency relationship to estimate time since MRCA (in days) as follows:

$$t = \frac{8\varphi}{5(1+\varphi)}\left(\frac{\lambda^*}{\varepsilon N_B} - \frac{1-\varphi}{\varphi^2}\right), \quad [3]$$

where N_B is the sequence length in base pairs and

$$\varphi = \sqrt{1+\frac{8}{R_0}}. \quad [4]$$

Furthermore, the fraction of identical sequences expected at that time is:

$$\mathrm{Exp}\left[-\varepsilon N_B\left(\frac{5}{8}t\,\frac{1+\varphi}{\varphi} + \frac{1-\varphi}{\varphi^2}\right) + O(\varepsilon^2 N_B)\right]. \quad [5]$$

The change in the Poisson distribution over time illustrates the increasing diversity expected under the model (Fig. S8). It is apparent that as time increases, the number of identical sequences decreases, and the frequency distribution of the intersequence HDs at various times after infection shifts to higher HD values.

Bayesian Analysis. The time, in days, to the MRCA for each patient was also estimated by using a Bayesian Markov Chain Monte Carlo (MCMC) approach, implemented in BEAST v1.4.1 (14, 15). The mean substitution rate was fixed at 2.16×10^{-5} substitutions per site per generation, and all analyses were carried out by using the general time reversible (GTR) substitution model with invariant sites and gamma-distributed rate heterogeneity (four gamma categories). The substitution and rate heterogeneity models were unlinked across codon positions, and we assumed exponential population growth and a relaxed (uncorrelated exponential) molecular clock. This model was used for analysis of the viral sequence alignment of each patient, and the MCMC algorithm was run for at least 10^7 (14) generations (logging every 1,000 generations; burn in was set to 10% of the original chain length), with additional runs carried out if the effective sample size for the estimate was <100. The results were visualized in TRACER (16). We repeated this analysis with the five free parameters of the GTR model fixed at values estimated by using the combined data from all acute patients inferred to be infected with a single viral strain by using the HyPhy package (17) and with alternative demographic and evolutionary models (relaxed uncorrelated molecular clock with logistic population growth and strict molecular clock with exponential population growth). Estimates and confidence intervals for the MRCA times were similar for the alternative relaxed clock models but ≈25% lower according to a strict molecular clock (data not shown).

Hypermutated Samples. Enrichment for APOBEC3G/F mutations violates the assumption of constant mutation rate across positions, because the editing performed by these enzymes are base- and context-sensitive. Enrichment for mutations with APOBEC3G/F signatures was assessed by using Hypermut 2.0 (www.hiv.lanl.gov), which compares each sequence in the sample to the consensus sequences. Hypermut detects an enrichment for G→A mutations that occur in the context of the APOBEC3G/F signature pattern, where the G is followed by either G or A, then by a base that is not C (in International Union of Pure and Applied Chemistry code, the pattern GRD where the first G changes to A). A contingency table is constructed, which is then used to obtain Fisher exact P values that test whether the extent of hypermutation is more than would be expected by chance. Single sequences that yielded a P value of ≤0.05 were considered significantly hypermutated and, therefore, were not included in other analyses; there were six such single sequences in the 102-patient dataset. In seven other cases, APOBEC3G/F patterns of substitution were enriched throughout the patient sequence set, but no one sequence was significantly hypermutated. In these cases, rather than testing each sequence separately, we assessed whether or not the entire sample was enriched overall for APOBEC3G/F signature mutations by collapsing all observed mutations into one sequence that represented all within-patient mutations, and then we used the Hypermut 2.0 program to compare this artificially constructed sequence to the consensus. Removing the APOBEC3G/F-mediated mutations in each of the 13 patients that showed enrichment for hypermutation showed that these samples otherwise conformed to our evolutionary model under no selection.

Viral Recombination. Twenty-four subjects of 102 studied were found to have been infected by two or more viruses. *Highlighter* tracings suggested that 16 of these subjects had HIV-1 *env* sequences with clear evidence of viral recombination. This interpretation was confirmed by statistical analyses with GARD (18) or Recco (19) recombination identification tools.

Replicative Fitness and Virus Outgrowth. It is possible that either immediately after transmission or in the first several of rounds of replication, a fitter form of the virus evolves and eventually grows out to become a dominant strain that we then detect. We argue that because of the short duration of time before plasma sampling, the long generation time of the virus, the low substitution rate in *env*, and the low R_0, this scenario is unlikely. Even if advantageous nucleotide (amino acid) replacements were to occur, their frequency based on random mutation throughout the 2.6-kb *env* gene would be expected to be exceedingly small, because most mutations are neutral or deleterious. Moreover, such viral mutants would be unlikely to replace the prevalent viruses unless the fitness advantage was large. For example, descendants of a virus at 20% replicative disadvantage compared to another still have more than a 5% chance of occurring in a sample of size 20, 10 generations (or roughly 20 days) later. Similarly, given the expected rate of one mutation every 20 replications, and an R_0 of 6, one expects the first mutation to occur after two generations of exponential growth. For such a form to be the only one, with 95% confidence, to leave descendants sampled in a sample of size 20 at 20 days later, it needs to have a reproductive advantage of more than 70%. The likelihood of any single random nucleotide substitution in a gene the size of *env* to confer such a selective advantage in anything other than a rare individual is remote and cannot plausibly explain the common finding of low diversity *env* lineages observed in 98 of 102 subjects in the present study.

Comprehensive Summary of Observed Evolution Patterns, Timing Estimates, and Statistics Comparing Results from a Model of Random Evolution in Acute Infection Under No Selection and Those Obtained By Using a Bayesian Method Implemented in BEAST. Dataset S1, Dataset S2, Dataset S3, and Dataset S4 contain demographic and sample information corresponding to each study subject along with summaries of model parameters and statistics, including the estimated time from the MRCA of the sequences found in each sample. The confidence intervals in the table do not take into account uncertainty in the viral mutation rate or the possibility of selection against a proportion of the nonsynonymous mutations. In each section of the table, the estimates of the time to the MRCA obtained by using BEAST were derived

from all of the sequence data, whereas the estimates based on the random evolution model were based on all sequences or edited sequence sets (see below). For the low diversity patients that conform to a random evolution model (Dataset S1), the estimated days from the MRCA are generally comparable to Bayesian estimates by using BEAST. In these cases, both models suggest the observed diversity profile could be expected to have emerged within a reasonable estimate of the time since infection given the Fiebig stage of the sample, and, thus, the infection scenario is compatible with evolution from a single transmitted variant. Samples with high levels of APOBEC3G/F-mediated mutations are summarized in Dataset S2. Again, the timing estimates according to the two models are roughly comparable when all substitutions are considered but with BEAST giving higher estimates. Substitutions with APOBEC3G/F signatures were then excluded for a second analysis by using the Poisson/Star model; when this was done, the observed sequence variation became compatible with this model and the estimated days to the MRCA was reduced. For patients with low diversity that significantly deviated from this model (Dataset S3), we noted in the table whether the best explanation for the deviation was transmission of multiple closely related strains, selection, or early stochastic events. Finally, for the 21 subjects with high diversity and multiple infecting strains (Dataset S4), the timing estimates from a random evolution model are based on times to the MRCA restricted to sequences from the dominant clade in the subject, because this provided a strategy to test whether each such clade plausibly represented the outgrowth of a single transmitted virus, which they did. Attempts to apply this model to the full sample dataset in these cases would clearly seriously violate the model assumptions, so we did not attempt this analysis on the high diversity samples. We did, however, model the minimum time required to achieve the maximum HD between sequences in the sample, as discussed in the main text, and this analysis is also summarized for each high diversity patient in Dataset S4; in all 21 cases, this analysis indicates that the MRCA existed in the donor before transmission and multiple viruses were transmitted. The BEAST estimates were derived from all of the data and indicated that the MRCA long preceded the date of infection, as expected for infection with multiple strains. This provided further corroborative support for the conclusion that multiple strains were transmitted based on the maximum HD described above. The finding that the Bayesian estimates of the time to the MRCA were consistently higher than the simulations based on the random evolution model can be interpreted as follows. The rate of early diversification of the virus depends on the time profile of production of virus from an infected cell, and not only on the average generation time and the number of cells newly infected in each generation. In this work, we have not investigated the parameters of the virus production profiles that govern viral diversification but merely set a weighted arithmetic average generation time as 2 days (13), both in the Bayesian model and in the simulations presented. The two, however, assume different profiles, and we expect this to lead to differences in the calculated rate of viral diversification. As an additional cautionary note, estimates based on both models neglect the effect of recombination although this is expected to have minimal effect, particularly for estimates based on highly homogeneous sequences exhibiting star-like phylogenies.

Power Studies. To better understand our likelihood of missing infrequent transmitted variants, we did a power study to explore the probability of sampling limitations. We show that with a sample of at least $n = 20$ plasma vRNA sequences (which was the case for 77 of 102 subjects), we could be 95% confident that a given missed variant comprised <15% of the virus population (Fig. S9). For 36 samples, for which $n \geq 30$, we could be 95% confident not to have missed any variant that comprised at least 10% of the total viral population. For 25 samples in which the number of sequences available was between 10 and 20, we had at least an 80% chance to detect a variant represented in ≥15% of the population.

***Env* Gene Cloning.** SGA-derived amplicons containing full-length *env* genes were molecularly cloned for protein expression and biological analysis. Transmitted/founder *envs* were identified as described along with SGA-derived *envs* from chronically infected clade B control subjects. The primers used to amplify these genes had been designed such that the DNA amplicon would contain a complete *rev*/*env* cassette. To reduce the probability of generating molecular *env* clones with *Taq* polymerase errors, we reamplified from the first-round PCR product under the same nested PCR conditions but used 10 fewer cycles. Correctly sized amplicons identified by gel electrophoresis were gel-purified by using the QIAquick gel purification kit according to the recommendations of the manufacturer (Qiagen), ligated into the pcDNA3.1 Directional Topo vector (Invitrogen Life Technologies), and transformed into TOP10 competent bacteria. Bacteria were plated on LB agar plates supplemented with 100 μg/ml ampicillin and cultured overnight at 30°C. Single colonies were selected and grown overnight in liquid LB broth at 30°C with 225 rpm shaking followed by plasmid isolation. Each molecular clone was sequence-confirmed to be identical to the transmitted *env* sequence(s) for each patient.

Env Phenotypic Analysis. The ability of cloned *env* genes to express functional glycoproteins was assessed as previously described by using an HIV-1 *env*-minus vector cotransfected into 293T cells to generate Env pseudotyped virions (20). These pseudovirions were then tested for entry into human JC53BL-13 cells (National Institutes of Health AIDS Research and Reference Reagent Program catalog no. 8129, TZM-bl), a HeLa-derived line that has been genetically modified so as to constitutively express CD4, CCR5, and CXCR4. This virus infectivity assay has been used extensively in the analysis of HIV-1 Envs and anti-HIV-1 neutralizing antibodies (20–22). JC53BL-13 cells also contain integrated luciferase and β-gal genes under tight regulatory control of an HIV-1 LTR, and, thus, virus entry can be quantitatively assessed over a broad range (23). JC53BL-13 cells (7×10^3) were plated in 96-well tissue culture plates (Falcon) and cultured overnight in DMEM supplemented with 10% FCS. For analysis of Env function, pseudovirions were quantified by p24Ag or RT activities and assessed directly for infectivity. For analysis of virus neutralization, 3,000 infectious units of virus were combined in a total volume of 60 μl with or without a 2× concentration of sCD4 in DMEM with 6% FCS and 80 μg/ml DEAE-dextran. After 1 h at 37°C, an equal volume of test or control plasma (10% vol/vol in DMEM plus 6% FCS or 5-fold dilutions thereof), mAb, fusion inhibitor, or chemokine coreceptor inhibitor was added. Monoclonal antibodies, as described (2), were kindly provided by the following individuals: Dennis Burton provided b12 and 2G12; Michael Zwick and Dennis Burton provided Z13e1; Herman Katinger provided 2F5 and 4E10; Susan Zolla-Pazner provided 447-52D; Lisa Cavacini provided F425-B4e8; James Robinson provided 17b; and David Montefiori provided HIVIG. The following reagents were obtained commercially: soluble CD4 (514-CD; R&D Systems); T1249 (Triangle Pharmaceuticals); and anti-CD4 mAb (555344; BD PharMingen). The coreceptor inhibitors TAK779 and AMD3100 were obtained from the National Institutes of Health AIDS Research and Reference Reagent Program (4983 and 8128). With the addition of ligand or antibody, this brought the final concentration of DEAE-dextran to 40 μg/ml. When sCD4 was used to trigger a conformation change in gp120 before cell attachment (24), the concentration was chosen so that the final 1× concentration after the addition of test antibody corre-

sponded to the IC_{50} of sCD4 specific for each virus. The virus plus sCD4 plus test antibody mixture was incubated for 1 h at 37°C. Media were removed entirely from the adherent JC53BL-13 monolayer just before the addition of the virus plus sCD4 plus test antibody to it. Cells were incubated at 37°C for 2 days and then analyzed for luciferase expression. Controls included cells exposed to no virus and to virus pretreated with normal human plasma (NHP) or control mAbs only. Relative infectivity was calculated by dividing the number of luciferase units at each dilution of test plasma or mAbs by values in wells containing NHP but no test plasma or mAbs. Neutralization was assessed by IC_{50} determined by linear regression by using a least-squares method. All samples were tested in duplicate, and all experiments were repeated at least three times to ensure reproducibility.

Some of the Envs that were characterized were sampled from the same patient. If an acutely infected patient was infected by one virus, then only that transmitted Env protein was used for assessing the phenotype. If an acutely infected patient was infected by more than one virus, however, each of the separately transmitted Envs was evaluated for phenotype. For the chronic controls, we again used the within-patient phylogenetic tree as a guide and selected one to four Envs that were dispersed throughout the tree to assess by phenotypic analysis. Thus, the Envs analyzed were not all independent with regard to the individuals from whom they were isolated, although each of the transmitted viruses represented an independent transmission event and each of the chronic samples represented a separate sublineage within the individual. Sometimes samples from the same individual had very similar patterns of sensitivity to the reagents tested, although they often had very distinctive patterns (see Fig. S7). For an initial test of the data, all points were treated as independent and subjected to a nonparametric Wilcoxon rank sum test to compare the distributions. Twenty comparisons were done; thus, uncorrected P values of <0.0025 were required to withstand a correction for multiple tests. To address the violation of the assumption of independence in the data, we then created 100 datasets, in which one Env per person was selected randomly from each patient for inclusion (transmitted Env, $n = 45$ individuals; chronic Env, $n = 13$ individuals). The median Wilcoxon P values when comparing these 100 datasets were <0.10 for the following three cases: T1249, $P = 0.027$; 4E10, $P = 0.034$; and 2F5, $P = 0.070$.

1. Jones NA, *et al*. (2004) Determinants of human immunodeficiency virus type 1 escape from the primary $CD8^+$ cytotoxic T lymphocyte response. *J Exp Med* 200:1243–1256.
2. Binley JM, *et al*. (2004) Comprehensive cross-clade neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. *J Virol* 78:13232–13252.
3. Cleghorn FR, *et al*. (2000) A distinctive clade B HIV type 1 is heterosexually transmitted in Trinidad and Tobago. *Proc Natl Acad Sci USA* 97:10532–10537.
4. Fiebig EW, *et al*. (2003) Dynamics of HIV viremia and antibody seroconversion in plasma donors: Implications for diagnosis and staging of primary HIV infection. *AIDS* 17:1871–1879.
5. Gaines H, von Sydow M, Pehrson PO, Lundbegh P (1988) Clinical picture of primary HIV infection presenting as a glandular-fever-like illness. *Br Med J* 297:1363–1368.
6. Clark SJ, *et al*. (1991) High titers of cytopathic virus in plasma of patients with symptomatic primary HIV-1 infection. *N Engl J Med* 324:954–960.
7. Schacker T, Collier AC, Hughes J, Shea T, Corey L (1996) Clinical and epidemiologic features of primary HIV infection. *Ann Intern Med* 125:257–264.
8. Little SJ, McLean AR, Spina CA, Richman DD, Havlir DV (1999) Viral dynamics of acute HIV-1 infection. *J Exp Med* 190:841–850.
9. Lindback S, *et al*. (2000) Viral dynamics in primary HIV-1 infection. Karolinska Institutet Primary HIV Infection Study Group. *AIDS* 14:2283–2291.
10. Lindback S, *et al*. (2000) Diagnosis of primary HIV-1 infection and duration of follow-up after HIV exposure. Karolinska Institute Primary HIV Infection Study Group. *AIDS* 14:2333–2339.
11. Markowitz M, *et al*. (2003) A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. *J Virol* 77:5037–5038.
12. Stafford MA, *et al*. (2000) Modeling plasma virus concentration during primary HIV infection. *J Theor Biol* 203:285–301.
13. Mansky LM, Temin HM (1995) Lower *in vivo* mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* 69:5087–5094.
14. Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.
15. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.
16. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 22:1185–1192.
17. Kosakovsky Pond SL, Frost SD, Muse SV (2005) HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.
18. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* 23:1891–1901.
19. Maydt J, Lengauer T (2006) Recco: Recombination analysis using cost optimization. *Bioinformatics* 22:1064–1071.
20. Wei X, *et al*. (2003) Antibody neutralization and escape by HIV-1. *Nature* 422:307–312.
21. Li M, *et al*. (2005) Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *J Virol* 79:10108–10125.
22. Mascola JR, *et al*. (2005) Recommendations for the design and use of standard virus panels to assess neutralizing antibody responses elicited by candidate human immunodeficiency virus type 1 vaccines. *J Virol* 79:10103–10107.
23. Wei X, *et al*. (2002) Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy. *Antimicrob Agents Chemother* 46:1896–1905.
24. Decker JM, *et al*. (2005) Antigenic conservation and immunogenicity of the HIV coreceptor binding site. *J Exp Med* 201:1407–1419.

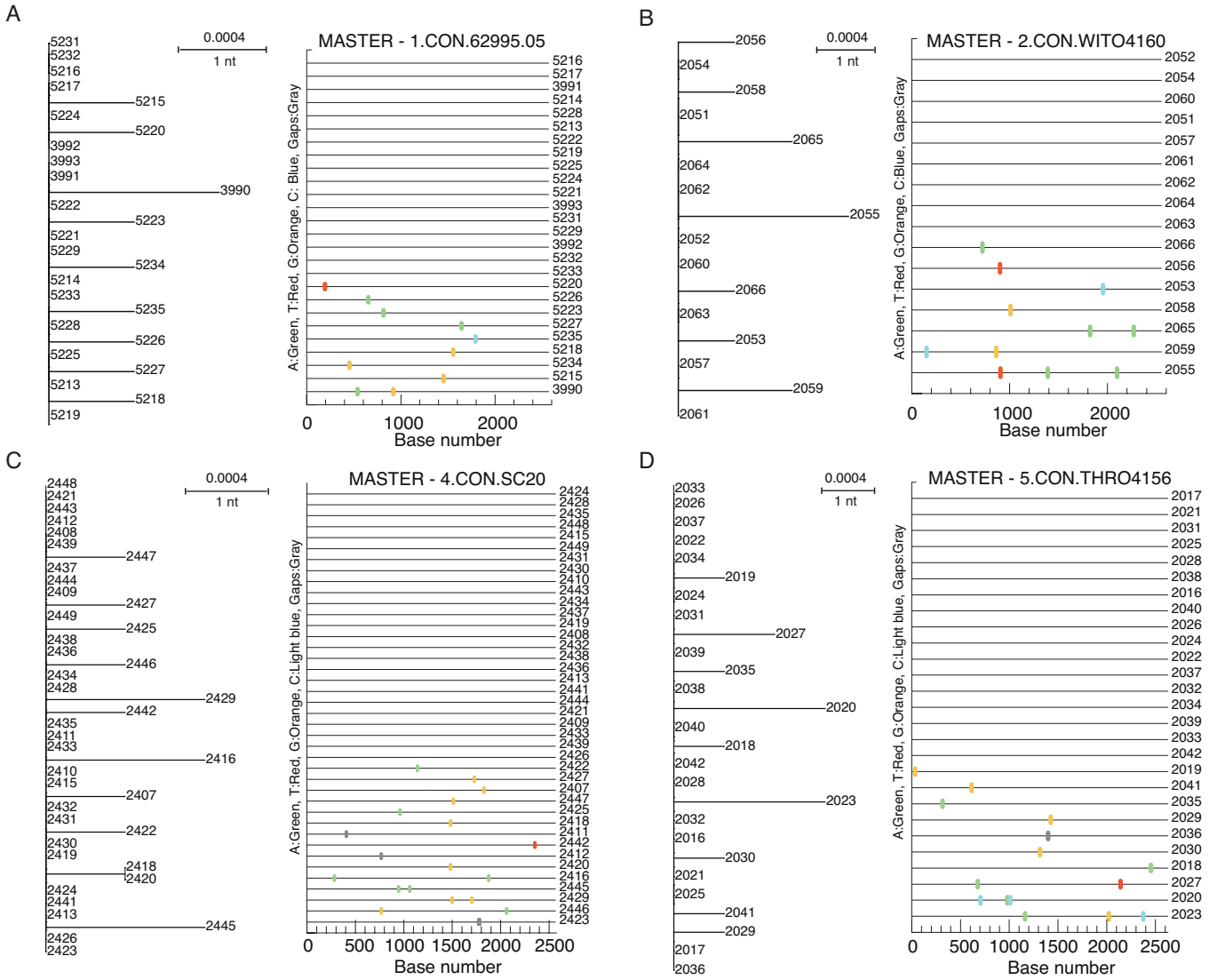

Fig. S1. Neighbor-joining and *Highlighter* analyses of *env* sequences from four subjects infected by single viruses. (*A*) Subject 62995 sampled at Fiebig stage I. (*B*) Subject WITO4160 sampled at Fiebig stage II. (*C*) Subject SC20 sampled at Fiebig stage IV. (*D*) Subject THRO4156 sampled at Fiebig stage V.

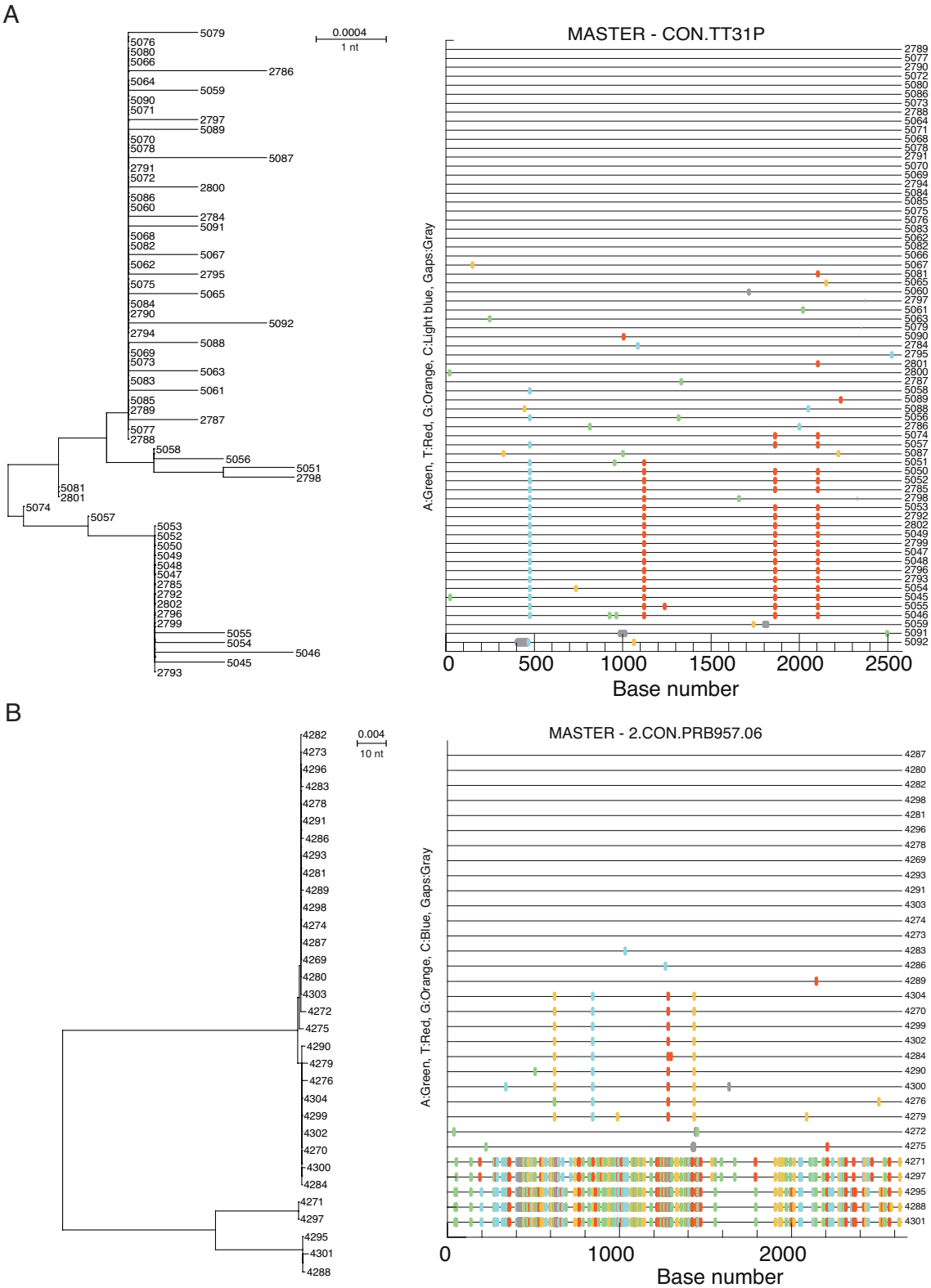

Fig. S2. Neighbor-joining and *Highlighter* analyses of *env* sequences from two subjects infected by closely related viruses. (*A*) Subject TT31 shows evidence of infection by two closely related viruses that are distinguished by a set of four nucleotide polymorphisms. In addition, there are interlineage recombinants evident. (*B*) Subject PRB957 shows evidence of infection by four viruses, including two lineages that differ from each other by a set of four nucleotides and two others that differ by ≈175 nucleotides.

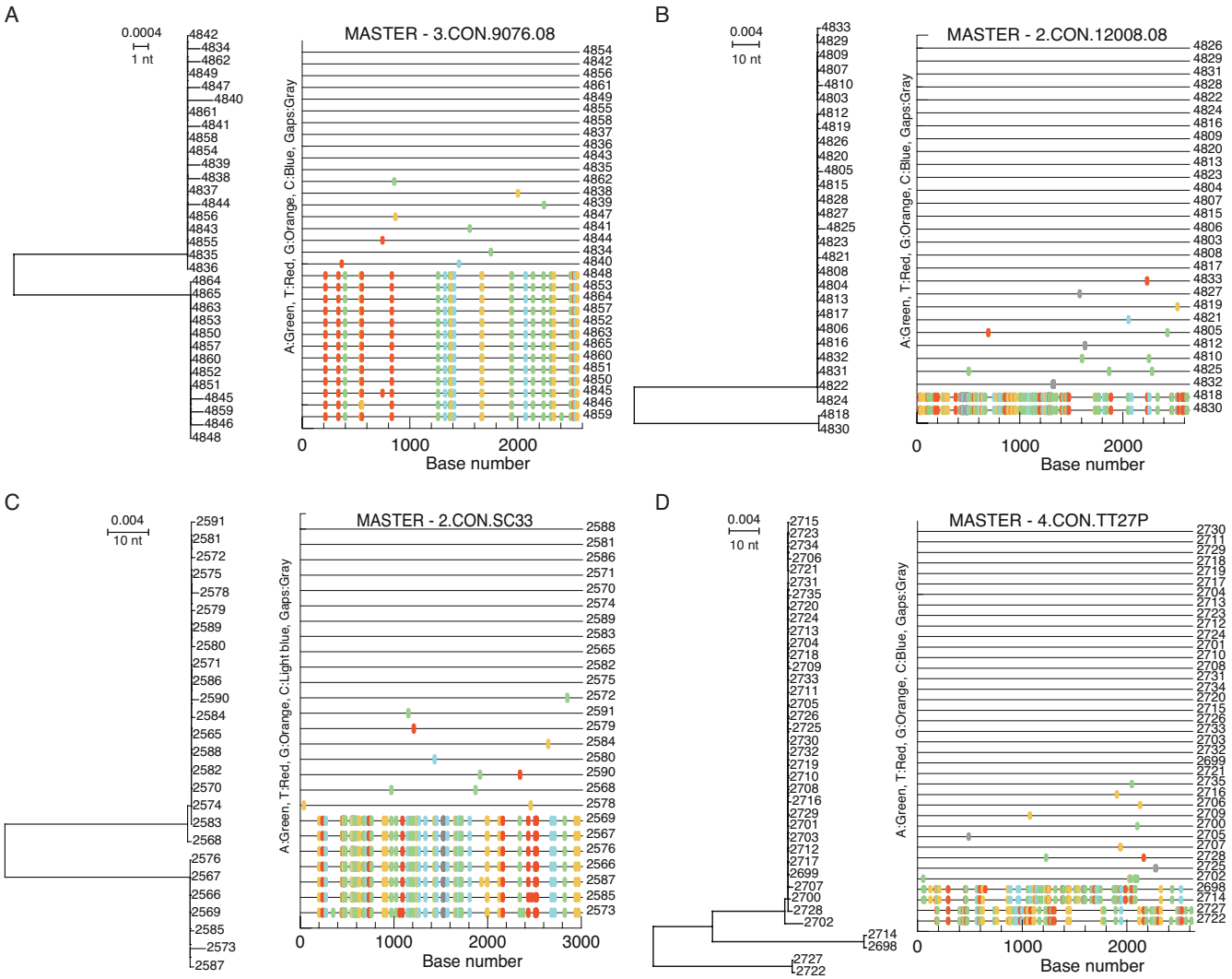

Fig. S3. Neighbor-joining and *Highlighter* analyses of *env* sequences from four subjects infected by more than one virus. Subject 9076 (*A*), 12008 (*B*), and SC33 (*C*) each was sampled at Fiebig stage II and demonstrates productive infection by two viruses differing in *env* by as much as 6% (*B*). Subject TT27P (*D*) was sampled at Fiebig stage IV and demonstrates infection by three viruses differing in *env* by as much as 5%.

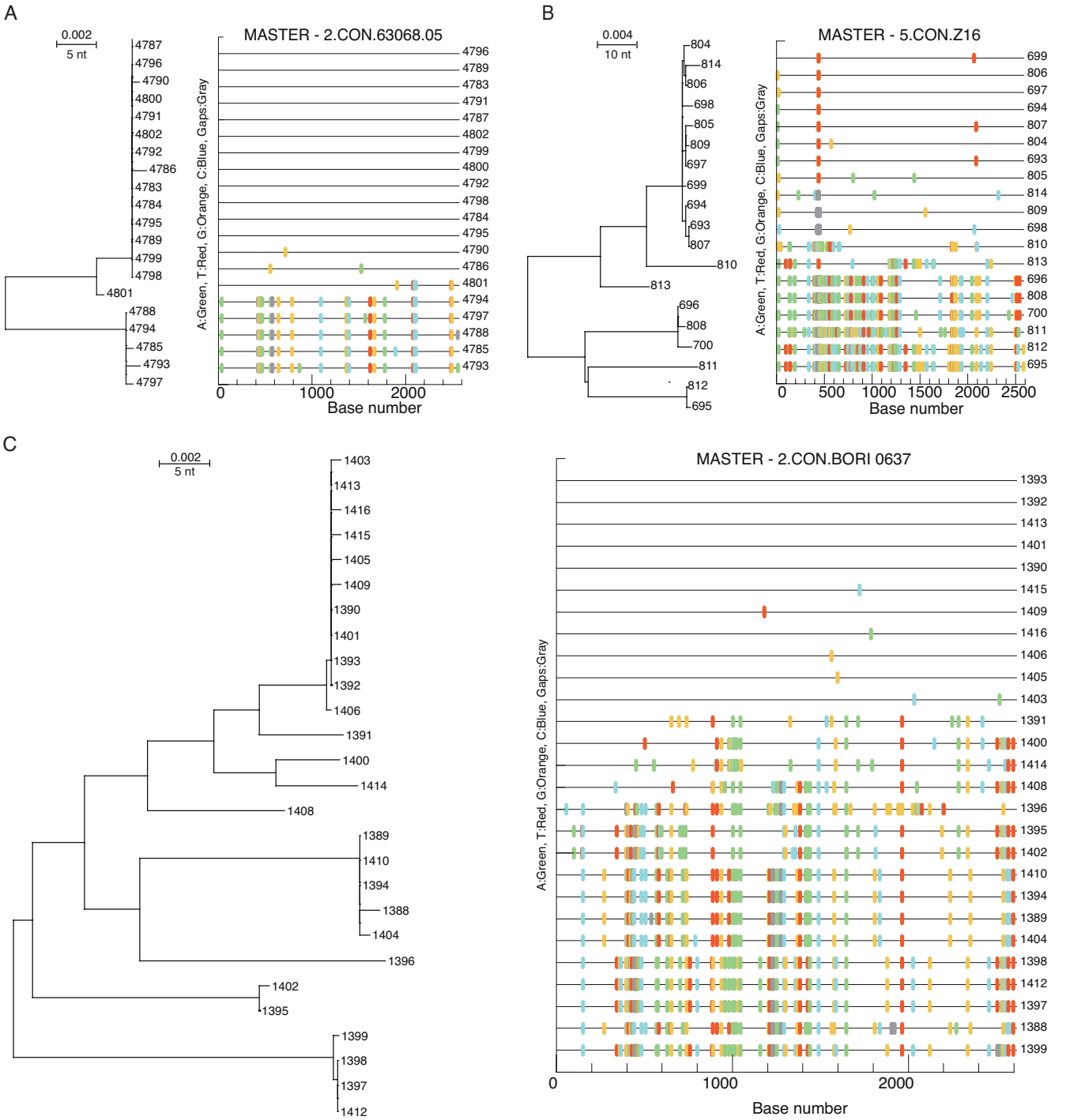

Fig. S4. Neighbor-joining and *Highlighter* analyses of *env* diversity in three subjects infected by more than one variant with evidence of recombination occurring during the acute infection period. (*A*) Subject 63068 sampled at Fiebig stage II shows productive infection by two viruses with recombination in sequence 4801. (*B*) Subject Z16 sampled at Fiebig stage V shows infection by four viruses with recombination in sequence 810 and 813. (*C*) Subject BORI0637 sampled at Fiebig stage II shows productive infection by five variants with evidence of recombination in sequence 1391, 1400, 1414, and 1408.

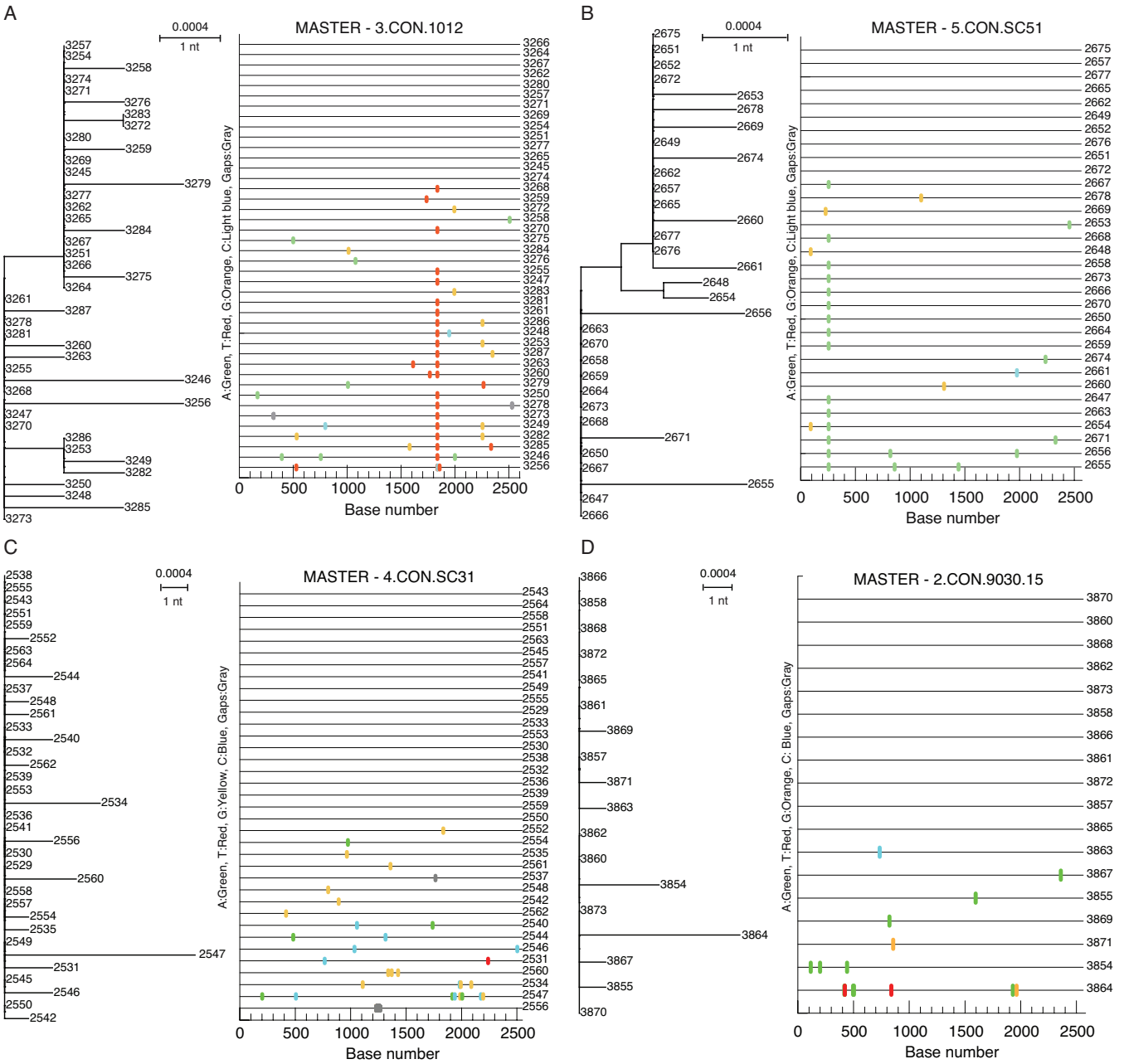

Fig. S5. Neighbor-joining and *Highlighter* analyses of *env* diversity in four subjects infected by single viruses whose evolution does not conform to model predictions. Subject 1012 (*A*) sampled at Fiebig stage III and SC51 (*B*) sampled at Fiebig stage V each shows evidence of an early stochastic change in *env* that was retained. Subject SC31 (*C*) sampled at Fiebig stage IV and subject 9030 (*D*) sampled at Fiebig stage II show evidence for non-Poisson distribution attributable to a single sequence with greater than expected diversity (sequences 2547 and 3864, respectively).

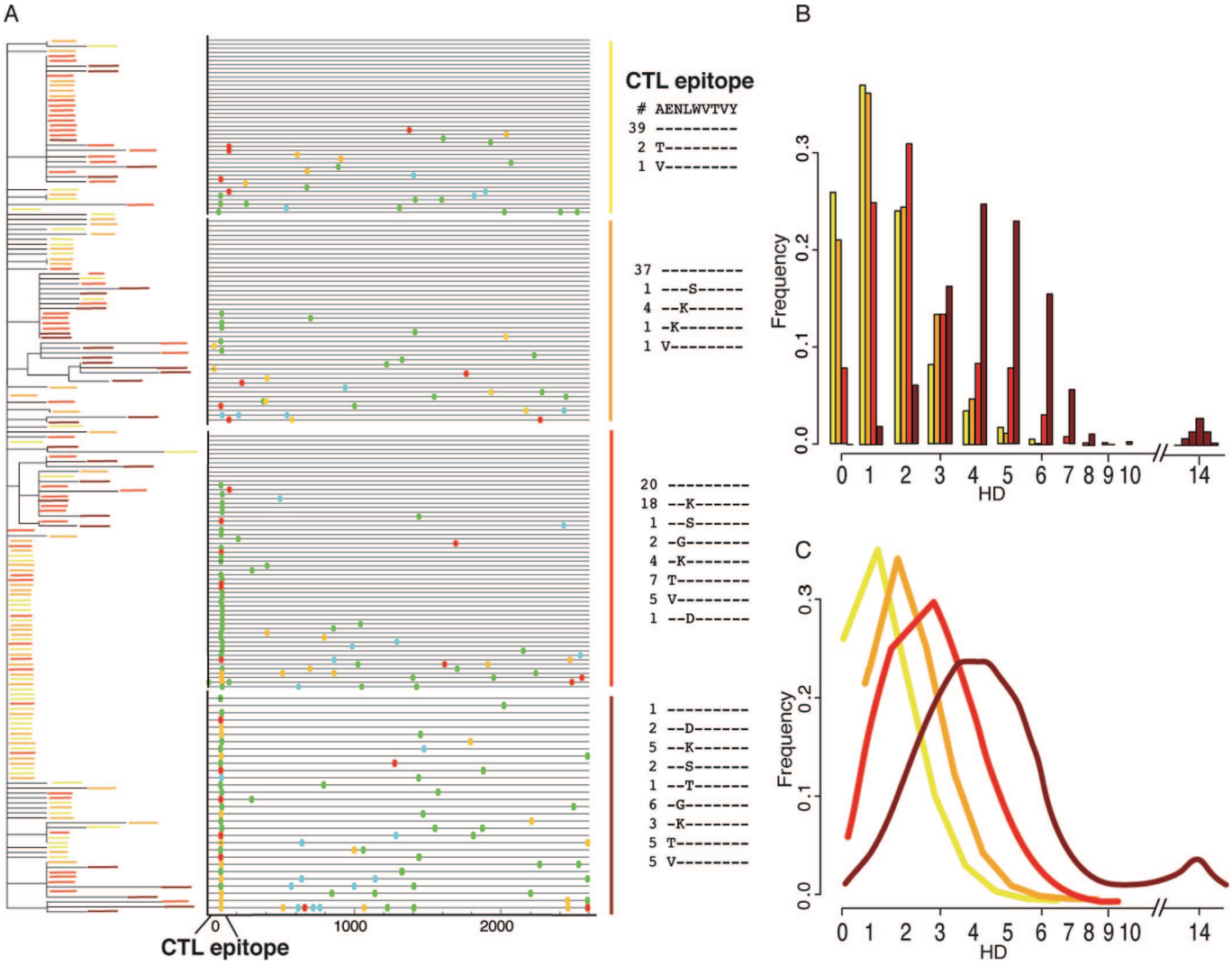

Fig. S6. Rapid evolution of CTL escape mutations. Sequences from the patient WEAU0575 were available at four time points early in infection, beginning in Fiebig stage II (indicated in yellow), and then 8 (Fiebig stage IV; orange), 15 (Fiebig stage IV; red), and 29 (Fiebig stage V; brown) days later. The neighbor-joining tree (*A*) shown is color-coded to depict the time point of each sequence, and the consensus of the first time point is at the root of the tree. Moving from left to right across the tree, later time points dominate the longer branch lengths, and clades begin to emerge, in part driven by mutations that concentrate in a CTL epitope that is indicated in the adjacent *Highlighter* plot. This epitope was recognized by homologous CTLs (1), and a variety of phenotypically proven escape mutations accumulated quickly as shown to the right of the *Highlighter* plot. Here, an alignment of the epitope is shown, with the different substitutions, and their frequencies are indicated. The observed HD distribution and its shift over time is shown (*B*) and compared with the expected distribution (*C*); the model clearly tracks the observed accumulation of diversity. The decline in the proportion of identical WEAU0575 sequences is depicted in Fig. 1 *A* by filled circles. (Note: a single hypermutated sequence was removed from the third time point.)

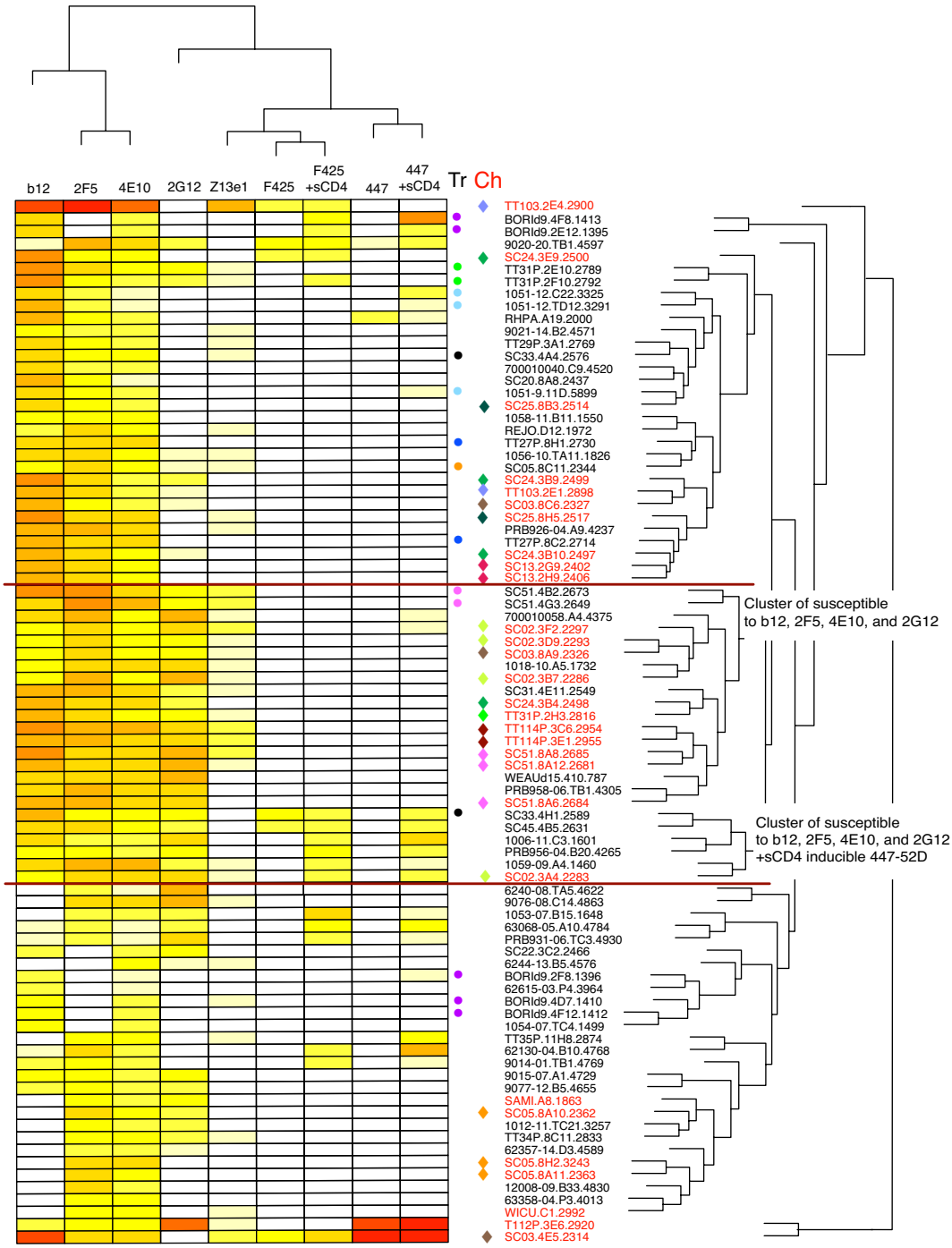

Fig. S7. Heat map dendrogram of neutralization results. The figure illustrates neutralization activity for each mAb against each transmitted and chronic Env tested. The colors in the boxes range from white, where even at the highest concentration tested neutralization was not observed, to light yellow, with red indicating increasing levels of neutralization sensitivity. Both patient data and antibody data are organized by agglomerative hierarchical clustering based on all pairwise Euclidian distances of log-transformed data to allow visualization of the data (2). The net result is that the Envs with the similar reactivity patterns are clustered together, and the antibodies with the most similar reactivity patterns are clustered together. The corresponding dendrograms are given along the top for the antibodies and on the right for the Envs. The Envs sampled during chronic infection are labeled red, with a Ch to indicate chronic, and the transmitted Envs are labeled Tr in black. The colored dots and diamonds preceding the subject IDs are color-coded to indicate when they are derived from the same subject. Several interesting patterns emerge: First, whereas, in some cases, Envs from the same individual display very similar neutralizing antibody sensitivities (e.g., subject TT31), in other instances, very distinctive behaviors are observed (e.g., subject BORI0637). Secondly, all Envs tested are susceptible to neutralization by at least one mAb. Thirdly, the Envs that are most susceptible to all four of the most broadly neutralizing mAbs known to date are clustered near the center of the figure; those that are susceptible only to b12, 2F5, 4E10, and 2G12 tend to be chronic Envs, whereas those that also have sCD4-induced susceptibility to V3 antibodies tend to be transmitted Envs.

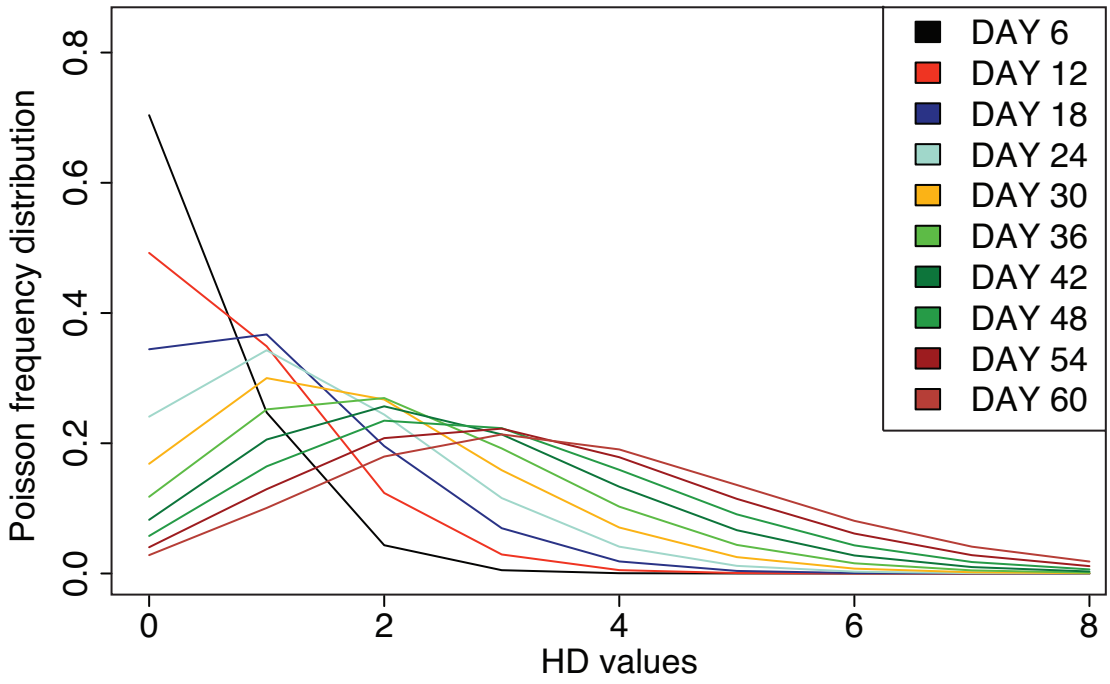

Fig. S8. Poisson Frequency Distribution. The graph shows the change in the Poisson distribution over time, to illustrate the increasing diversity expected under the model of random virus evolution. As time increases, accumulation of mutations shifts the distribution to higher maximum Hamming distance (HD) values. See *SI Text*.

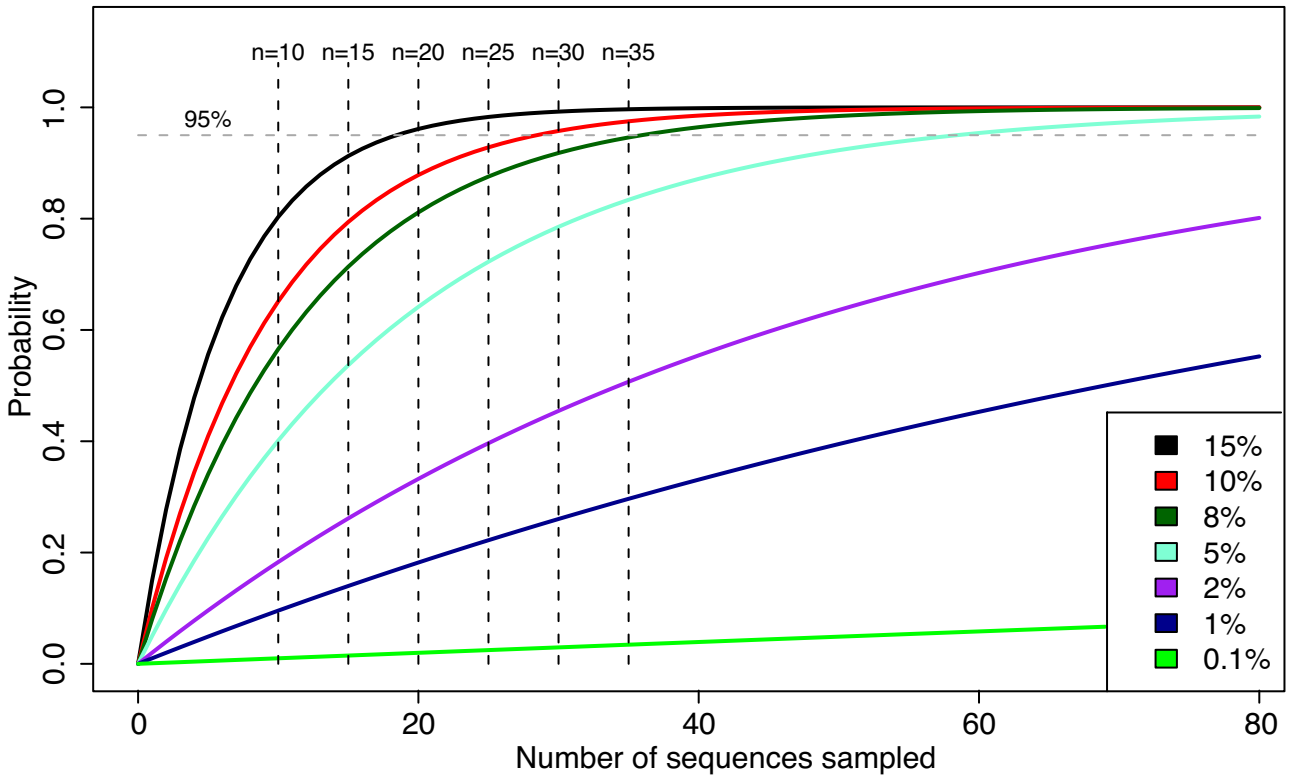

Fig. S9. Power analysis to estimate the likelihood of detecting infrequent transmitted variants. See *SI Text*.

Other Supporting Information Files

Dataset S1 (XLS)
Dataset S2 (XLS)
Dataset S3 (XLS)
Dataset S4 (XLS)
Dataset S5 (XLS)
Dataset S6 (XLS)
Dataset S7 (XLS)